

Analysis of Demonstration Data for Redistricting and Voting Rights Act Use Cases: Production Settings

August 10, 2021

Coordinator:

Welcome and thank you for standing by for today's conference, All participants will be in listen-only most of the duration of today's call. I would now like to turn the conference over to Michael Hawes. Sir, you may begin.

Michael Hawes:

Thank you operator. Good afternoon everyone and welcome to the latest in our Webinar series on understanding the 2020 Census Disclosure Avoidance System. I'm Michael Hawes, Senior Advisor for Data Access and Privacy in the Census Bureau's Research and Methodology Directorate.

Today we're going to be providing an analysis of the Disclosure Avoidance Systems production settings for the Redistricting and Voting Rights Act use cases. I'm joined today by a number of my colleagues including James Whitehorne, Tommy Wright, Kyle Irimata, Jen Shopkorn and Shelly Hedrick who will be answering your questions during the presentation and helping to moderate additional questions and discussion at the end of the Webinar.

If you would like to submit a question during the presentation today please use the Q&A feature of the WebEx platform and please be sure to send your questions to all panelists so that the entire team can see them and the best person can respond to it.

Next slide please. Before I begin I would like to acknowledge my Census Bureau colleagues within the agency and our collaborators who have contributed to the information that's

contained in today's presentation. I would also like to acknowledge and sincerely thank our many external stakeholders and the data user community in general for the ongoing feedback that they've provided which has been invaluable in our ongoing efforts to design and improve the 2020 Census Disclosure Avoidance System. And finally, I just want to state that any opinions and viewpoints expressed today are entirely my own and do not represent the opinions or viewpoints of the US Census Bureau.

Next slide please. So those who have attended our prior Webinars in this series will recall that on June 9 we published a press release informing the public that our Data Stewardship and Executive Policy Committee had set the production settings for the 2020 Census Public Law 94-171 Redistricting Data Summary Files. Those settings are essentially the implementation of the parameters and privacy-loss budget for use in protecting the confidentiality of the 2020 Census redistricting data.

Next slide please. In setting those parameters and in setting the privacy-loss budget we relied quite extensively on the extensive feedback that we've received from many in the stakeholder community, particularly the feedback that we received on the April 2021 demonstration data set that we released where we took 2010 Census data and ran it through our Disclosure Avoidance System and published the resulting data so that data users could compare the privacy-protected data using the Disclosure Avoidance System to the published 2010 tabulations.

The extensive feedback that we received from those demonstration data covered a number of themes. Some of the major topics that were included in that feedback included concern about accuracy for American Indian and Alaska Native tribal areas and other what we call off-spine geographies which I'll talk a little more about later.

And also concerned about accuracy for places, for minor civil divisions and tract level data. Some lingering concerns about geographic or characteristic bias that was still present in those demonstration data, concerns about the overall accuracy of race and ethnicity statistics and about occupancy rates for households.

Next slide please. Based on that feedback and our own extensive internal analyses of accuracy and fitness for use measures, the Data Stewardship Executive Policy Committee made a number of improvements and changes to the parameters that are going to be used for the actual production run of the 2020 Census P.L. 94-171 redistricting data. Those final parameters include a larger privacy-loss budget, PLB is the acronym, within epsilon of 17.14 for the persons level file and epsilon of 2.47 for the units file.

We also included improvements to our optimized geographic post-processing hierarchy that the TopDown Algorithm utilizes. We allocated additional privacy-loss budget to total population counts at a number of geographic levels, as well as additional privacy-loss budget to race and ethnicity statistics and occupancy rates particularly at the block group level and above.

Next slide please. If you want to learn more about how those parameters and settings were implemented you can see the full spreadsheet of privacy-loss budget allocations across the query sets and across geographic levels via this link [here](#). And these slides will be available on our Web site later so you can just click the link directly.

Next slide please. Most notably the improvements and changes in the production settings reflect that increase in privacy-loss budget which I mentioned. On this table comparing the parameters of the April 2020 demonstration data to the production settings you can see, for example, for the person's level file the Privacy-loss budget increased from an epsilon of 10.3 in the April demonstration data to 17.14 in the actual production settings, because this is a

logarithmic scale that is a significant increase in privacy-loss budget which will result in substantially greater accuracy.

You can also see that with that added privacy-loss budget we also reallocated shares of that privacy-loss budget across the geographic levels with substantial increases in privacy-loss budget allocations at the state, county tract, and optimized block group levels. And this will significantly increase accuracy for those against the accuracy targets that our Data Stewardship Executive Policy Committee set.

Next slide please. You can also see based on this -- and I apologize that it's a little small -- that we also made changes to how that privacy-loss budget was allocated for inquiries that the Disclosure Avoidance Systems TopDown Algorithm utilizes in performing the privacy protections. In particular we allocated substantial additional privacy-loss budget to the total population query at the state, county and tract levels, and to the Hispanic by race queries to improve the accuracy of race and ethnicity statistics at the national, state, county, tract, and block group levels as well.

Next slide please. On August 12, so later this week, the Census Bureau will be releasing the official 2020 Census Public Law 94-171 redistricting data files. We will also be releasing these in an easier to use format by September 30 of this year, September 30, 2021.

Next slide please. In parallel with the release of the official 2020 Census redistricting data the Census Bureau will also be releasing the final set of 2010 Census demonstration data reflecting the final production settings used by the Disclosure Avoidance System to protect the 2020 Census data. So these will be 2010 Census data run through the Disclosure Avoidance System with the exact same settings and privacy-loss budget that were used to produce the privacy protected 2020 Census redistricting data.

These are the same data, these data that we'll be releasing in parallel, are the same data that were used to produce the production settings detailed summary metrics files that we released on July 1 of this year. And if you want to see those you can click on this link here.

Next slide please. In today's presentation we're going to look at two sets of analysis of the 2010 Census demonstration data using those production settings for the Redistricting Voting Rights Act to use cases. The first of these is an empirical study of two aspects of the TopDown Algorithms output for redistricting looking at reliability and variability.

This is an updated version of the Wright and Irimata 2021 study for which we had a prior Webinar in this series. The second analysis that we'll be presenting is an analysis of the impact of these production settings on the ability to identify and assess majority minority districts at a variety of district sizes.

Next slide please. Both sets of analysis that we're going to be discussing today are going to involve comparisons of the 2010 Census demonstration data, those protected with the differentially private 2020 Census Disclosure Avoidance System to the published 2010 Census tabulations. Now it's important to note that the 2010 Census used a different form of noise confusion known as data swapping whereby households records were swapped across geographies in order to protect the confidentiality of the members of those households.

For the 2010 Census data the number of individuals in the household and the number of voting age individuals in each household were held invariant, so no noise was added into total population or total voting age population across geographies. But individual characteristics, particularly those for race and Hispanic origin, were swapped across geographies.

So the 20210 published tabulation did include noise in the race and Hispanic origin tabulations against which we're comparing. So differences between the 2010 demonstration data

protected using the Disclosure Avoidance System and the published 2010 Census tabulations that are presented in the following analysis reflect both the noise from the 2020 Census Disclosure Avoidance System and the impact of the 2010 Census' swapping methodology on characteristics data.

Now I do want to note that the internal tuning that the Census Bureau was doing, the tuning that was done to essentially determine the final production settings, was not done in comparison to the swapped data. Our internal analyses that we used to tune the algorithms for accuracy and fitness for use were actually done in comparison to the raw census edited file, the unswapped data.

And as has become somewhat apparent in some of the recent analyses the differences between those two can be substantial. So be aware as you're looking at the results of these analyses that because we're going to be comparing to the swapped published tabulations that much of the error that you're going to see can be attributed to the 2010 swapping as well as to the 2020 Disclosure Avoidance System.

Next slide please. So before we get into the specific analyses I do want to look briefly at average error and total population counts by geographic level. Now remember I did say a moment ago that the total population counts and the total voting age population counts for the published 2010 Census data did not have any noise added. For the 2020 Census there will be noise in the total population and total voting age population counts that are published.

This table here taken from our production settings Detailed Summary Metrics, which we released last month, shows the average amount of error in the total population counts by different levels of geography. So for example at the county level your average county will see an error due to privacy protection of about plus or minus about 1.75 people.

Your average minor civil division will see an error of plus or minus 2.7 people. Average incorporated place an average error of 3.5 in total population, tracts plus or minus 1.9, urban blocks plus or minus 4 and rural blocks plus or minus 1.6 six people. So this this gives you an idea of the amount of noise due to privacy protection that you can see in those total population counts.

Next slide please. All right, so the first of the analyses we're going to look at today, as I mentioned, is an updated version of the Wright and Irimata 2021 study that we covered in a prior Webinar. And this is an update based on the new production settings demonstration data. This is empirical study of two aspects of the TopDown Algorithm output for redistricting, reliability and variability. And that paper is available on the Census Bureau's Web site.

Next slide please. So the Wright and Irimata 2021 paper asked two fundamental research questions. The first question is, what is the minimum total population of a district to have reliable characteristics of various demographic groups for redistricting purposes?

Now because districts cannot be defined in advance of the redistricting process this portion of the Wright and Irimata 2021 analysis uses two existing geographic levels as proxies for districts to assess accuracy. The first are block groups which are largely on the geographic processing hierarchy, and the other are places and minor civil divisions, those that are off of the geographic processing hierarchy. The second research question that Wright and Irimata 2021 asked is how variable are data that are protected using the 2020 Census Disclosure Avoidance System for districts in Rhode Island and for three additional jurisdictions?

Next slide please. So part one of the right Wright and Irimata analysis relies on a key measure known as the Difference of Ratios, or DR, for the largest demographic group in any particular geography. To calculate the Difference of Ratios, DR, they take the absolute value of the difference between the ratio of the largest demographic group as a proportion of the total

population for that geography in the published swapped 2010 Census data. And that group's ratio, again, as a proportion of the total population, in the data protected using the 2020 Census disclosure Avoidance system's TopDown Algorithm.

In their formulas here you can distinguish between the two sets of data by the subscripts, SWA, short for swapped, meaning the published 2010 tabulations, and TDA, short for TopDown Algorithm, the data protected using the Disclosure Avoidance System. When looking at the difference of ratios small values of DR imply that the results across the two sets of data are close. Larger values of DR indicate greater divergence of values across the two modes of privacy protections.

Wright and Irimata assert that if the DR is sufficiently small, less than 5% for the purposes of this analysis, when you're comparing across the swapped and TDA versions of the data at a particular level of geography, then Wright and Irimata consider the TDA version to provide a reliable characteristic for that geography.

Next slide please. In this table taken from the paper, you can see how the DR is calculated for various demographic groups within a particular block group. In this case the block group of my colleagues at the Census Bureau with 1560 residents. In this example there were 133 Hispanic individuals in the published swapped 2010 tabulations and 141 in the data produced with the Disclosure Avoidance System's production settings. Taking those as ratios of the total population for the block group either 1560 people or 1598 persons for the respective data sets, you can then calculate the difference of ratios as 0.003.

Next slide please. From these data you could also calculate the degree of error in total population by taking the absolute value of the difference in counts and dividing by the total population in the swapped data. As you can see the error for this particular block group is 2.4% for the total population counts and 2.6% for the voting age population counts. Both of which

are less than the 5% accuracy target that was used when tuning the accuracy at the largest demographic groups.

This is a feature of the way our DAS tuning was done. By tuning for the accuracy of the largest groups as a proportion of their total populations we were also tuning the total populations for those areas and achieving greater accuracy in the process.

Next slide please. So when you look at the difference of ratios for the largest demographic groups in each block group across all block groups Wright and Irimata wanted to identify the overall size of block groups necessary to achieve that 5% accuracy target at least 95% of the time. And as you can see in this table accuracy improves as block group size increases. And that accuracy threshold was crossed once block groups reach the total population size of 450 to 499 individuals.

Next slide please. And as you can see from this excerpt from farther down the table accuracy continues to increase as block groups increase in size, both in terms of the percentage of block groups meeting the 5% threshold, which is the right-hand most column, but also in terms of the percentage of block groups that meet the tighter 3% and 1% threshold the second and third columns from the right respectively. As those block groups get larger those targets get tighter and the proportion of block groups that meet those targets increases.

Next slide please. The size of block groups necessary to achieve this degree of accuracy was also consistent across different runs of our algorithm. Wright and Irimata reran their analysis on 25 independent runs of the 2010 Census data through our Disclosure Avoidance System. And this table shows the block group size where that 5% threshold was achieved along with the percentage of block groups meeting that target for each of those 25 independent runs.

Next slide please. Now those of you who listened in on prior Webinars know that the TopDown Algorithm processes data along the central geographic hierarchy or geographic spine and that

accuracy can vary whether a geography is on the spine or off the spine. So to assess the accuracy for geographic areas farther from that processing spine Wright and Irimata also performed their analysis on a number of off-spine entities of interest.

These included minor civil divisions such as boroughs or townships in strong and CD states and places including incorporated places and census designated places in weak and CD states. And as you can see in this table the accuracy target of 5% at least 95% of the time was achieved for places in MCDs with total populations of at least 200 to 249 persons with over 96% of those geographies meeting that accuracy target.

Next slide please. And once again you can see that the percentage of geography's meeting the target and the tightness of the resulting accuracy also continued to increase as minor civil divisions and places increase in size. For example over 99-1/2% of places and MCDs with 850 people or more meet that tighter 3% accuracy target the second to right-hand column.

Next slide please. And again these results are consistent across the 25 independent runs of the Disclosure Avoidance System with all runs meeting the threshold for places and MCDs with 200 or more individuals and several of the runs meeting it for those with populations of just 150 individuals or greater.

Next slide please. Now Wright and Irimata used block groups, places, and minor civil divisions as proxies for small voting districts. But many within the redistricting community are particularly interested in accuracy for congressional districts and state legislative districts at the upper and lower house levels.

Examining these districts, the smallest of which has a minimum population of 3,173, Wright and Irimata proceed to demonstrate that all 25 independent runs of the TopDown Algorithm using 2010 Census data meet the established accuracy target, as expressed in this difference of

ratios, 100% of the time for all districts with sizes of 3,150 to 3,199 persons representing the smallest of these legislative districts. And again accuracy of these results improves as district size increases.

Next slide please. In part two of their analysis, Wright and Irimata examined those 25 independent runs of the 2010 Census data through the 2020 Census DAS using the same production settings for their variability against a range of redistricting use cases, including Rhode Island two congressional districts, their 38 upper state legislative districts and their 75 lower state legislative districts.

They also examined them against three jurisdictions supplied to us by the Department of Justice, including Panola County Mississippi has 2,180 blocks, Tate County School District Mississippi that has 784, and Tylertown or Walthall County, Mississippi with 136. And the idea this was to examine the variability across districts of different sizes.

Next slide please. Now the specific measure that Wright and Irimata used to assess this variability was the average relative variation among the population over all of the demographic groups. And the specific formula they used for this can be seen here in definition four. This measure across all of the demographic groups in a district, across all of the independent runs of the Disclosure Avoidance System, can be thought of as the overall coefficient of variation.

Next slide please. And looking at the districts and jurisdictions analyzed once again we can see that the smallest districts do exhibit greater variability across runs. But as district or jurisdiction size increases that variability decreases substantially.

Next slide please. So Wright and Irimata 2021 has two primary empirical conclusions. The first is their message on reliability. And they state that for any block group with a total population count between 450 and 499 or larger and for minor civil divisions and places with between 200

people and 249 people or larger, the difference between the TopDown Algorithm's ratio of the largest demographic group and the corresponding published 2010 Census tabulation using swapping ratio for the largest demographic group, is less than or equal to 5 percentage points at least 95% of the time.

And no congressional or state legislative district fails this test meaning that these districts have the 5% criterion holds 100% of the time. And their key empirical message on variability is that the relative variability in the TopDown Algorithm decreases as we consider larger pieces of geography and population. At a high level their analysis tends to show less relative variability using the 2020 Census redistricting data production settings than the April 2021 demonstration data that we previously released.

Next slide please. So next, we're going to turn to an analysis of the impact of the production settings on the identification of majority minority districts. So to perform this analysis we examined 436 congressional district, 1,946 state upper legislative districts and 4,785 state lower legislative districts.

Next slide please. You need to go one more. The demographics that we examined were a variety of racial groups tabulated differently across the different tabulations included in the redistricting data and those that were included in SF1 following the 2010 Census.

The total population by race categories from the redistricting T1 table or the SF1 P8 table included white alone, black alone, American Indian Alaska Native alone, Asian alone, some of the race alone and black and black plus white. The redistricting P2 table or the SF1 P9 tables that provide breakdowns by total Hispanic or not Hispanic by race, that included Hispanic, not Hispanic white alone and so on.

The redistricting P3 or SF1 P10 tables which provide voting age population by race were analyzed across these various groups. And then the P4 P11 tables that look at Hispanic or not Hispanic voting age population by race examined the following list of categories as well.

Next slide please. So what we wanted to find was, were there are cases of districts at the congressional, state upper legislative or state lower legislative district where a district would have been identified as majority minority in the published 2010 tabulations, but would not be seen as having a majority in the protected-with-the-production settings or vice versa, and we did find examples where flips occurred in both directions.

For the white alone population, total population, we saw one state upper district move from 50.01% to 49.99% white alone. And we saw one congressional district move from 49.99% to 50.01% in the production settings data. For black alone we saw one similar shift from 50.08% black to 49.95% black in terms of total population for a state lower district.

Next slide please. For total Hispanic we saw one shift in a state lower district from 49.92% in the published 2010 tabulations to 50.02% in the production settings of the Disclosure Avoidance System.

Next slide please. Not Hispanic by race we saw a few shifts as well. We saw for the not Hispanic white alone population one shift from 50.02% percent not Hispanic white, to what came just underneath 50.00% rounded to two significant digits, it became 50. We also saw one state upper district from 49.95% to 50.02%.

We saw one not Hispanic black alone state lower district move from forty 49.91% to 50.05%. And we saw one state lower district with a not Hispanic American Indian or Alaskan Native population alone shift from 50.1% to 49.37% in the resulting data.

Next slide please. For voting age by race we saw a few shifts here as well. Voting age white alone population we saw one shift from 50.01% to 49.95% at a state lower district. For the voting age population black alone we saw one shift negative and one shift positive, one for the upper and one for the lower. And for voting age population black and black and white we saw two districts move from not majorities to majorities, one state lower district and one congressional district. Again, both very tight margins here.

Next slide please. For Hispanic voting age population, the P4/P11 tables, we saw two districts move from a very tight margin of 50.02% to 49.99% or 49.97% one was a state upper one was a state lower.

Next slide please. For the not Hispanic voting age population by race again we saw one not Hispanic white district to state upper moved from 50.1% to 50.0% with rounding fell just below that threshold. We saw four district that did not have majorities of not Hispanic black alone in the published tabulations just eked out over the 50% threshold in the 2020 production settings data. Two, excuse me, three state lower districts and one state upper district.

Next slide please. And for the not Hispanic black and black plus white voting age population we saw five shifts, two state lower districts that went from just over 50% to just under 50.06% to 49.95% and 50.03% to 49.95%. And then we saw two state lowers and one congressional district where they just eked out over the 50% threshold as well.

So what conclusions can we draw from this analysis? Well comparing the production settings 2010 demonstration data to the published 2010 Census tabulations, the data identified 25 districts out of the 7,167 that we analyzed from those congressional districts, plus the upper legislative, plus state legislative getting a 0.3% of all of those districts where a demographic group could be considered to flip from majority to minority or vice versa across the two sets of data releases.

So this occurred in both directions, 11 groups went from majority to minority and 14 went from minority to majority. It's important to note that no flips involved both a racial or ethnic group's total population and their voting age population. That is districts drawn to have majorities of both total population and voting age population would be more stable across these analysis.

And all flips involve the very, in some case a very, very small number of individuals, in districts that were very tightly drawn, usually within a few hundredths of a percent of the 50% mark using the published 2010 Census tabulations. And if you remember our discussion about the impact of the swapping algorithms used for the 2010 published tabulations where race and ethnicity did have noise injected, they were swapped across geographies, that's a level of precision that would have been greatly impacted by the noise injected from that 2010 swapping algorithm.

Next slide please. So if you'd like to learn more about our work on our Disclosure Avoidance System and stay informed of new and upcoming updates please subscribe to our newsletter. You can just search disclosure avoidance at [census.gov](https://www.census.gov).

Next slide. And please visit our Web site. We've got a wealth of information about our reasons for adopting our modernization of disclosure avoidance and our efforts to design and improve our 2020 Disclosure Avoidance System. We have fact sheets, issue briefs, frequently asked questions, videos and more all available on our Web site. Again just search disclosure avoidance at [census.gov](https://www.census.gov).

Next slide. And we do have a new video protecting privacy and Census Bureau statistics which serves as a great introduction to the type of noise infusion that we're using for the 2020 Census data. You can find it on our YouTube page and on our Disclosure Avoidance page.

And next slide. And with that I will turn things over to my colleague, Jennifer Shopkorn, who will moderate some questions and answers. Thank you Jen.

Jennifer Shopkorn:

Thank you Michael. That was a great presentation. We've been trying to keep up with the questions in the chat that we've been getting but we'll take the remaining time to try and discuss some of those topics we've seen there verbally here so folks get the benefit of our analysis and our answers.

Just want to remind folks if you do ask a question in the Q&A box in WebEx platform please address it to all panelists so we can make sure the entire team can take as many questions as possible. Michael, I'm going to start with a question here while some folks are answering some more in the chat for you.

One questioner, one participant in the Webinar wanted to get a little bit more of a sense about whether the Census Bureau has looked at the effects on local redistricting at smaller levels of geography?

Michael Hawes:

So I actually will defer to some of my other colleagues as well. We've got James Whitehorne and Tommy Wright and Kyle Irimata. I think all of us have been involved in various of the analysis. James or Tommy or Kyle so you want to speak up?

Tommy Wright:

Can the question be repeated again?

Jennifer Shopkorn:

Of course.

Tommy Wright:

Please.

Jennifer Shopkorn:

Yes, absolutely. The question is, "Has the Census Bureau looked at the effects on local redistricting at smaller levels of geography?"

Tommy Wright:

In our analysis -- this is Tommy Wright -- in our analysis we - I think the lowest is the block group. So what we've looked at Michael has done an excellent job of sharing. I yield to James.

James Whitehorne:

Yes. This James Whitehorne from the Redistricting and Voting Rights Data Office at Census. So in the analysis that Tommy has done we've looked at, there's three cases that are demonstrated in the paper that's been written and provided, did look at some other use cases that were involved there.

We had our consultations with a lot of folks in the redistricting community and with our colleagues over in the voting section at the Department of Justice to really try to understand the use case around redistricting and that the concerns for accuracy really come in for this small area geography. And so that's why we focused on tuning towards not just the block groups but places and minor civil divisions because those tend to be areas that more represent what a redistricting plan would look like because they're considered off spine geography.

So by doing that and working to get that accuracy target to I believe it was 200 to 249 -- so one of my colleagues are correct me if I'm wrong for...

Tommy Wright:

Yes.

James Whitehorne:

...and MCDs. That was driving our search to try to ensure that the data were fit for use for small area redistricting as well.

Jennifer Shopkorn:

Thanks James.

Tommy Wright:

I will point out that some of those block groups are very small. In fact I think the very smallest one had 82 people in the entire block group which was to be divided into four districts. So I just make that as a comment.

Jennifer Shopkorn:

Thank you both. We have another question here. Michael, hoping we can visit one point from earlier but someone had a little bit of confusion I'm hoping you can clear that up. "In the demonstration comparisons how many districts went from majority to minority for AIAN?"

Michael Hawes:

Excellent question. And I can actually I'm not driving the slide. So Shelly if you can move us back to... it's the ...keep going one. Yes okay there we go.

So there was one AIAN district that did shift. It was the not Hispanic - it was when measured as not Hispanic, American Indian Alaska Native alone out of total population, so the P2 tables.

There was one district that went from 50.1% to 49.37%. And again if you look at the numbers it was a difference of approximately 63 people in that particular case. But again as I mentioned

before we did not see the same shift if you looked at American Indian Alaska Native voting age population. So this was one.

And again with all of these, it's important to remember that these were all subject, these are all comparisons against the published data that included the noise from swapping. So how much of these shifts are attributable just to swapping is a question that cannot be shared publicly.

Jennifer Shopkorn:

Thank you Michael. Another question here in the Q&A box that we're getting is a little bit of clarification. "What was the group black plus black and white was used - that was to in part of the analysis?" Could you revisit that topic as well please.

Michael Hawes:

So that was the distinction that we were asked to include in our office by our colleagues at the Department of Justice. I'll actually ask James if he can speak more to when and how that is used.

James Whitehorne:

Well I mean I -- this is James again -- and I won't interpret how it's used because that's really subjective based on the end user. But it was something that was requested as we were told by our colleague that it was a category that they sometimes look at. I can't characterize how they actually use it though.

Jennifer Shopkorn:

Thank you both. Hopefully that was helpful for the individual who asked that question.

James Whitehorne:

Sorry something just occurred to me Jen in regards to that question as well. And I remembered that in the Office of Management and Budget, guidance on how to aggregate race for the purposes of Civil Rights and Voting Rights Enforcement, it's Bulletin 00-02, they do have the black plus black and white, as one of the categories that they suggest should be aggregated together for that analysis.

Jennifer Shopkorn:

Thanks for that follow-up James. Appreciate it. Michael, I'm going to take this next one to you. Sorry my computer just did something funny. Hopefully it'll come back in just a second all of these work from home tech challenges. Okay, there we go. "Can you talk a little bit about whether users have access to the optimized block groups since that's not a traditional census geography?"

Michael Hawes:

So that's a great question but it is a tricky one. The optimization that's done essentially involves redrawing, for the purposes exclusively of the TopDown Algorithms processing, redrawing the boundaries of block groups so that they bring these off-spine geography's closer to the processing spine.

It does not impact how anything is tabulated or published. Those are still done with the traditional tabulation block groups. But in terms of the algorithm's processing it redefines blocks into these kind of optimized block groups to bring those off spine entities closer to what is being directly measured and processed via the algorithm.

And that process occurs as part of the algorithm running. So it's not like the optimized block groups are determined in advance. With each run of the algorithm the algorithm identifies the optimal way of essentially moving blocks around to bring these off spine entities closest to the

processing hierarchy. And also separating out the block groups that have group quarters by type so that they don't essentially bleed or diffuse population into their surrounding area.

So we don't have a list that we can provide because again those are determined when the algorithm runs and each run would do that slightly differently. However I can say this, the drawing of those was done to bring particular geographies closer to the spine.

So the optimization was done and to bring minor civil divisions census designated places, incorporated places and American Indian Alaska Native tribal areas closer to the spine. So the optimization was drawn to make those closer to these optimized block groups.

Jennifer Shopkorn:

Thank you. That's helpful I hope. It's certainly helpful for me to hear that explained a little bit more. We've gotten a question here I think is probably a good baseline to revisit for folks. Someone is asking if we could talk a little bit more about how the 5% reliability was chosen to measure the redistricting accuracy. Can you talk about that a little bit? And Michael if you're the right person that would be great if I have directed it incorrectly please let me know.

Michael Hawes:

So I can speak generally but then I'd love to defer to Tommy and Kyle as well. The way we were tuning the algorithm we needed quantifiable targets against which to tune. And I think it's no surprise that when discussing accuracy targets with many among the data user community, those targets have often been expressed in kind of abstract terms but we needed like quantifiable targets against which to tune.

And so we needed to pick some threshold to determine, some threshold against which to tune the allocation of privacy-loss budget to achieve. Tommy your paper really implemented that 5% rule. Do you want to speak a little bit about why you selected it?

Tommy Wright:

I think what you've said is pretty good. I will confess that this problem came to us from James Whitehorne and John Abowd. And we sort of formulated a question that we thought we could answer after several months of sort of data exploratory data analysis.

We not only have in the background in arriving at this 95 of the time and 5 percentage points we looked at, as you can see, a little bit of 1%, one percentage, .03. And we also looked at not just the largest group but also the two largest groups and also a little consideration to the three largest groups.

So the criterion - there are lots of criterion that people can define. But we thought that this was one that could be useful in conveying just what a minimum size a district could be in order to have - to sort of serve as a boundary point. If you have a district below this boundary point then you have less reliability and for demographic groups inside that district and if it's above that you tend to have more confidence.

So it's - the criterion actually just came out of the data itself. It just after several months of exploring the data. I will remind people it is an empirical result but it's based in data. But we've seen it over and over again in different types of data and we have begun to think a little bit about how we can formalize this phenomenon that we seem to see in the data.

((Crosstalk))

Tommy Wright:

I think the honest answer is really initially lots of exploratory data analysis. And Kyle do you - I don't know if my colleague Kyle wants to add a little bit to that. Maybe not.

Michael Hawes:

Okay, I would also add to that while this was one of the primary accuracy targets that we were using to tune for the Redistricting and Voting Rights Act use cases. This was by no means the only target that we were using. And in fact when our Data Stewardship Executive Policy Committee was doing their extensive evaluations that led to the setting of the production settings and parameters in that final overall privacy-loss budget and its allocation they looked at a wide range of accuracy measures when determining exactly what those production settings should be. This was just one of those.

Jennifer Shopkorn:

All right thank you. Just scrolling here, we got a lot of questions coming in and we have some colleagues answering them in the chat. So rest assured we're trying to get to all of your questions. I want to make sure I don't ask verbally one that we're answering on a chat as well. Michael, someone is asking for clarification about one of the slides we showed about a person level and housing unit PLB. They said they thought that we had an additional PLB on there. Can you please talk about a little bit more?

Michael Hawes:

Sure. That was one of the early slides Shelly if you want to scroll back to that. Keep going. One more I think. One more. There we go.

So the TopDown Algorithm, which is used for the production of the P.L. 94-171 Redistricting Data Summary Files and will also be used for the Demographic Housing and Characteristic Files which are the next set of data products that will be coming out. The TopDown Algorithm processes the person-level data separately from the household-level data.

And so those of you who have been using our demonstration data files over the past year, almost two years now, will have noticed that those demonstration data come in two files.

There's always PPMF persons file and a PPMF units file. And because they're run separately they need separate privacy-loss budget allocations.

Any time you perform any query on the confidential data in a formally private system you need to expend privacy-loss budget in the process and then allocate it across those queries within a data product. So for the production settings we assigned a privacy-loss budget of 17.14 for all of the tabulations that produced the person's-level file. In the context of the redistricting data product that would be essentially all of the data in tables P1 through P5.

And then in the units file, the housing units file, we expended a privacy-loss budget of 2.47. That's essentially all of the data that's in table H1 of the redistricting data product.

Jennifer Shopkorn:

Thank you Michael. All right, just scrolling through some questions here. Appreciate everyone's patience and time. Michael, can you talk about the - remind folks what's happening - at what population level the housing occupancy will be accurate?

Michael Hawes:

So I mean I think that's kind of an open question because I mean what is meant by accurate? Again there's different ways of assessing accuracy. There's different ways of assessing fitness-for-use for different uses.

One thing I will say I mean the underlying premise of the Disclosure Avoidance System, the underlying premise of the TopDown Algorithm, is to produce noisy block level data that can be aggregated effectively to produce reliable and accurate statistical results once you've aggregated them, essentially taking noisy pixels and creating a clear picture once you've grouped those pixels together.

So I think with any block level data I would always express a note of caution about relying on individual blocks for analysis. However, as you group those blocks together the accuracy and fitness-for-use will increase. So as you group blocks, housing counts for example, together into geographies of interest for example, neighborhoods or tracts, those statistics will get substantially more reliable. And the larger those areas get, the larger the underlying population, the more accurate they will be as well.

I will caution however that some data users, when looking at the demonstration data, have been essentially making calculations across the universes. So essentially trying to make calculations across the persons and households universe dividing number of persons by number of households. At low levels of geography that will yield some problematic results.

We have, as one of our later data products, we will actually be producing the Person Household Joins data, as we've called them, in our detailed demographic and housing characteristics files that will be released later on. And those are being processed together and don't suffer from the separation of the processing of persons and the processing for units that the TopDown Algorithm does.

So if you are interested in persons per household I would encourage you to only make that calculation at larger levels of geography for example the larger tracts or above if you want highly reliable data. Otherwise wait until the detailed demographic and housing characteristics that have those detailed household data included which will be coming out at a later date.

Jennifer Shopkorn:

Thank you. Going to ask the next question of James Whitehorne from our redistricting office. James, we've gotten a few questions in the chat asking about the type of data we're putting out on Thursday. If folks know on Thursday, this Thursday August 12, we're releasing our next round of 2020 Census results, this is our redistricting data.

And we have a news event that folks can tune into at 1 o'clock. That information is available on our Web site at census.gov. But James, I was hoping you could touch on a few of the questions we've gotten about the scope of the data that we're releasing, what that looks like in terms of the type of data ...

James Whitehorne:

Sure.

Jennifer Shopkorn:

...that we're putting out in those top line counts? Thanks.

James Whitehorne:

Yes sure, on the update. So there's going to be quite a quite a lot of information that's going to be coming out during the press announcement on Thursday, so not tomorrow but the day after, at 1 o'clock.

We'll start by publishing the actual summary files to our FTP sites so that the people who need those can go ahead and start downloading those and begin their processing. But there's going to be several reports, America Counts stories, that are going to be published. Those America Counts stories will have some of the information and talk about some of the trends that are being seen in the data.

There's also going to be some visualizations that are going to be quite powerful that will allow people to look at different sets of data. They'll have ranking tables at the state and county level so people can look in more depth at those top line numbers.

They'll be a mapping tool that's available that people will use, an interactive map. Certainly quite a few resources that by the time that press announcement is over that people will be able to go to and they'll be able to get those top line numbers and actually start to dig down into some of the detail as well.

Jennifer Shopkorn:

Thank you so much. I hope that is helpful for folks. And while we're talking about our big day on Thursday just wanted to remind folks that there have been a lot of blogs, and stories and data visualizations that we've already released on that topic just to help people understand the data before it comes out on Thursday.

All of that is available. There's a really easy one stop shop Web site where you can access all that information. Just go to census.gov/rdo, that is, R as in Roger, D as in David O. And you'll find a wealth of information already. And certainly on Thursday the volume of information will increase considerably.

It has currently a lot of communications material so data visualizations, videos, explanations about how to access the data, how to work in the data files for the legacy files that we're putting out on Thursday, blogs explaining the type of information we're putting out, how the census process works. It also has a social media tool kit if you're interested in spreading the word about 2020 Census data.

And after Thursday we'll have additional contextual information and it will also have links to access the FTP site and data. So I hope folks will visit that bookmark it and you'll be well prepared for Thursday and you'll I hope continue to visit that, we'll continue to put information out there.

All right, I know we're just about at the end of our time here. I'm going to take one quick scroll. It looks like most of the questions have been answered. Michael, I want to give you one more opportunity just to share any other parting thoughts with folks if there's anything that you wanted to follow-up on since our discussion has continued here?

Michael Hawes:

Sorry, I'm just in the process of scrolling through questions here. So no final thoughts on my end, sorry.

Jennifer Shopkorn:

No worries, no worries. And just as I'm doing the same and scrolling through questions. An important question has been asked here about the time the data will be released on Thursday. Our press conference begins at 1 o'clock Eastern.

You can find the information on that and the link for that. It is at [census.gov](https://www.census.gov). And if you go through the top navigation on our Web site to look at the menu there and you go into the newsroom you'll see the information posted there. Anyone can join and watch. It will be broadcast live and media is welcome to register and, credentialed media is welcome to register and ask questions.

And yes there will be just - questions keep popping up. You all have good ones. There will be lots of visualizations and maps on Thursday, many of them interactive if folks are familiar with Tableau. Our platform allows us to zoom in and out and look at data for specific levels of geography. So yes we'll have information available at lots of geographic levels.

All right, any other last thoughts Michael? I know we are right at time here? I want to be respectful...

Michael Hawes:

Just thanking everybody for...

Jennifer Shopkorn:

...of everyone's time.

Michael Hawes:

...participating today. And if you have additional questions definitely check out the other Webinars that we've done in the series if you haven't already. You can send questions into us via the email on our Web site. And we look forward to continuing this engagement as we move past the redistricting data towards the Demographic and Housing Characteristics Files.

Jennifer Shopkorn:

All right, thanks everyone. Take care and hopefully see you Thursday.

Coordinator:

Thank you. That does conclude today's conference. We do appreciate your attending. You may disconnect at this time.

END